

Distributed indexing: performance analysis of Solr, Terrier and Katta Information Retrievals

ABSTRACT

Information Retrieval (IR) systems are currently facing a continuous challenge due to the increasing size of datasets. Extremely, large data from different aspects is gathered each day resulting in huge increase in the scale of the raw data available across the internet. Indexing, as the main function of IR systems, is becoming a time-consuming problem. Therefore, efficient indexing of a large volume of data is now a critical requirement of modern IR systems. High performance indexing is performed nowadays over the use of MapReduce programming model. MapReduce is a programming paradigm that enables massive processing and large collections distribution across multiple (hundreds or thousands) commodity computers. To shed some light on this issue, this paper presents a detailed performance analysis of distributed indexing for Solr, Katta and Terrier with the context of MapReduce. In particular, this study compares and analyzes the distributed indexing performance of three frameworks using 1GB, 3GB, 6GB, and 9GB subsets of TREC dataset as the processing power increase. The experiments measure the indexing average time, then throughput, speedup, and efficiency of indexing process. The results show that, Terrier performance is the best in the presence of large collections and scalable processing power. While, Solr performance is the best when having limited computing power and small document collections. Finally, the experimental results show that, Katta produced the worst indexing average time among the three frameworks but its speedup scales linearly with processing power and collection size.

Keyword: Indexing; Information Retrieval; Hadoop; Katta; MapReduce; Solr; Terrier